B3clf: A Resampling-Integrated Machine Learning Framework to Predict Blood-Brain Barrier Permeability

Fanwang Meng,[†,‡,¶] Jitian Chen,[†,¶] Juan Samuel Collins-Ramirez,[†] and Paul W. Ayers[*,†]

[†]*Department of Chemistry and Chemical Biology, McMaster University, Hamilton, ON, Canada, L8S 4L8*

[‡]*Department of Chemistry, Queen's University, Kingston, ON, Canada, K7L 3N6*

[¶]*These authors contributed equally to this work.*

E-mail: ayers@mcmaster.ca

Abstract

Developing accurate, computationally efficient, and reliable predictive models for small molecules' blood-brain barrier (BBB) permeability is challenging due to the class imbalance often found in collections of reference data. We use resampling techniques to address class imbalance and build 24 types of machine learning models, which we developed using comprehensive hyperparameter optimizations. We evaluated our model against those from previous studies, which provides insight into optimal classification models and resampling techniques that are relevant beyond BBB permeability. In addition to classifying unknown compounds on the basis of BBB permeability, the predicted probabilities are provided to facilitate further improvements and comparative benchmarking, and to report the models' confidence in their predictions. To disseminate our findings, we developed `B3clf`, a highly efficient, user-friendly tool that facilitates BBB permeability prediction, which can be accessed as open-source software (https://github.com/theochem/B3clf) or as a web app (https://huggingface.co/spaces/QCDevs/b3clf). The newly curated external dataset for BBB is hosted at `https://github.com/theochem/B3DB`.

1 Introduction

The blood-brain barrier (BBB) maintains homeostasis in the central nervous system (CNS)[1,2] and protects the CNS by inhibiting the passage of toxins and pathogens from the blood[3]. However, its selective permeability also poses a challenge for the delivery of neuroactive molecules, such as drugs, into the CNS. It is estimated that approximately 100% of biomolecular pharmaceuticals (e.g., peptides and monoclonal antibodies) and 98% of small molecule drugs are unable to penetrate the BBB[4]. BBB permeability is also a key consideration when developing chimeric antigen receptor–modified T (CAR-T) cell-based therapy for brain tumors[5,6]. Therefore, understanding small molecules' BBB permeability is crucial for prioritizing promising candidates and avoiding investment in compounds unlikely to reach brain targets.

To overcome this challenge, various experimental approaches have been proposed to measure the penetration of BBB molecules *in vivo*[7–11]. These methods typically focus on two types of BBB permeability data: (1) the logarithmic ratio of molecular concentration in the brain to blood in the steady state ($\log BB$)[12]; and (2) the permeability surface area product ($\log PS$)[13]. However, experimental methods for assessing the BBB permeability of small molecules are expensive, time-consuming, labor-intensive, and low-throughput. Therefore, computational prediction of BBB permeability has become an attractive yet challenging problem in CNS drug discovery and development.

Various computational approaches have been proposed for predicting BBB permeability. The most reliable approaches use molecular dynamics (MD) simulations to directly simulate the solubility and/or transport of molecules across the BBB. An early MD study correlated the computed solvation free energy in water with $\log BB$[14]. More recent studies have employed steered MD[15], enhanced sampling techniques[13], and unbiased MD[16] to provide a detailed characterization of physical interactions between molecules and membranes and obtain atomic-level insights into molecules' (in)ability to cross the BBB. However, MD simulations require substantial computational resources and expertise to set up and analyze, which has motivated the development of cheaper and faster approaches. Data-driven methods are appealing because BBB permeability is primarily determined by molecules' physicochemical properties such as topological polar surface area (tPSA), number of hydrogen bond donors and acceptors, and pK_a, which can be quantitatively represented and learned from data[17,18]. This understanding has led to the development of quantitative structure-activity relationship (QSAR) for BBB permeability[19–23]. More recently, interest in machine learning (ML) approaches has grown. ML methods extend QSAR by incorporating "synthetic" structural properties (e.g., molecular fingerprints) and more sophisticated mathematical models. There are two types of ML models for BBB permeability predictions: (1) regression for $\log BB$ or $\log PS$; and (2) binary classification of molecules as BBB permeable (BBB+) or BBB impermeable (BBB-). Due to the limited size of publicly accessible datasets, there are relatively few ML models for the regression of $\log BB$ values. Current studies have used methods including multiple linear regression[12], neural networks[24–26], and support vector regression[27]. In contrast, a large and diverse range of ML models for the classification of molecular BBB permeability is available, including random forests (RF)[28,29], classification and regression trees (CART)[30], binomial partial least squares (binomial-PLS)[31], decision tree induction (DTI)[32], support vector machines[33], and generalized linear models (GLM)[34 35].

While more data is available for BBB classification than for regression, model generalizability to unseen data remains limited because the chemical space covered by the available data is insufficient to build a sensible decision boundary.[36] Consequently, QSAR and ML models are prone to overfitting. This explains the counterintuitive observation that decision-boundary-based algorithms, such as support vector machines, often perform better on smaller datasets (e.g., 1593 molecules) than larger datasets (e.g., 1990 molecules) for BBB classification[33]. Moreover, class imbalance in available BBB datasets also compromises model performance. The BBB prevents most chemicals from entering the brain, meaning that most molecules are BBB-. However, experimental datasets tend to overrepresent BBB+ samples, making them the majority class. This creates a rather unusual situation where the imbalance in nature (strongly skewed toward BBB-) and the scientific literature (somewhat skewed toward BBB+) are opposed.

Different methods have been proposed to correct for the biases that imbalanced data can introduce in ML classification models, including methods at the algorithmic level (e.g., biased minimax probability machine (BMPM)) and resampling strategies (e.g., synthetic minority oversampling technique (SMOTE)[37]). There are three main types of resampling strategies: (1) undersampling on the majority class; (2) oversampling over the minority class to generate synthetic data points; and (3) hybrid models[38,39]. Undersampling can help increase the sensitivity of a classifier to the minority class[37,40], but tends to ignore information from the majority class. In contrast, oversampling techniques can improve minority class representation without data loss. Among these, some of the most widely used methods belong to the SMOTE family. SMOTE generates new synthetic data points for each minority class instance by interpolating existing data points[37]. Inspired by the success of SMOTE, many variants have been proposed,[41] including Borderline SMOTE[42], k-means SMOTE[43], adaptive synthetic sampling (ADASYN)[44], and density-based SMOTE (DBSMOTE)[45]. Borderline SMOTE, an adaptive SMOTE variant, generates artificial data points near the decision boundary to improve the classification performance. This approach is based on the assumption that minority class examples located near the decision boundary are most susceptible to misclassification[42]. k-means SMOTE is designed to reduce noise in the data and better represent the decision boundary[43]. ADASYN generates synthetic minority class samples in an adaptive pattern by focusing on entries that are hard to learn, thus reducing the bias of the classifier[44].

Given the difficulty of experimental measurements, it is infeasible to address the data imbalance problem by generating additional minority class (BBB-) instances at high throughput. This motivates the use of resampling methods for BBB classification[46]. Wang *et al.* used support vector machines as the base algorithm and applied SMOTE to a dataset of 2358 molecules, giving a model with a specificity of 0.833[47]. Using the same dataset, a recurrent neural network (RNN) was proposed in combination with SMOTE, which achieved strong performance on the training data; however, this study did not utilize a training/testing protocol[48]. More recently, SMOTE was integrated with extreme gradient boosting (XGBoost)[49] and evaluated using a 0.75:0.25 train/test split on the same dataset[50]. Good precision and recall (sensitivity) with an F_1 score of 0.91 was achieved, but specificity was not reported. Another similar study using the same dataset applied SMOTE followed by a feed-forward artificial neural network (ANN) to generate feature vectors, which were then input into kernel principal component analysis (KPCA). This model achieved an overall accuracy of 97.11%, specificity of 98.42%, and sensitivity of 97.35% on the testing set[25].

In this study, we address the challenges of limited generalizability caused by small, imbalanced datasets by developing predictive models for BBB permeability using machine learning classification algorithms combined with resampling techniques. The training dataset was curated from 50 literature and publicly available sources, as described by Meng et al.[51], and is illustrated in Figure 1 (A). We employed four classification algorithms (decision trees, k-nearest neighbors (kNN), logistic regression, and XGBoost) in combination with six sampling strategies: SMOTE, Borderline SMOTE, k-means SMOTE, ADASYN, random undersampling, and a baseline with no resampling ("common"), as shown in Figure 1 (B). Model performance was systematically evaluated using a comprehensive set of metrics, including accuracy, sensitivity, specificity, precision, F_1, MCC, GEOM, BACC, ROC_AUC, and AP. The trained models were further benchmarked against state-of-the-art approaches using a newly curated external test set comprising of 175 compounds (171 BBB+ and 4 BBB-), provided as Supporting Information Table SI. 1. We also implemented a Hugging Face web server to improve reproducibility and make it easier for non-computational researchers to access BBB prediction models for drug discovery.

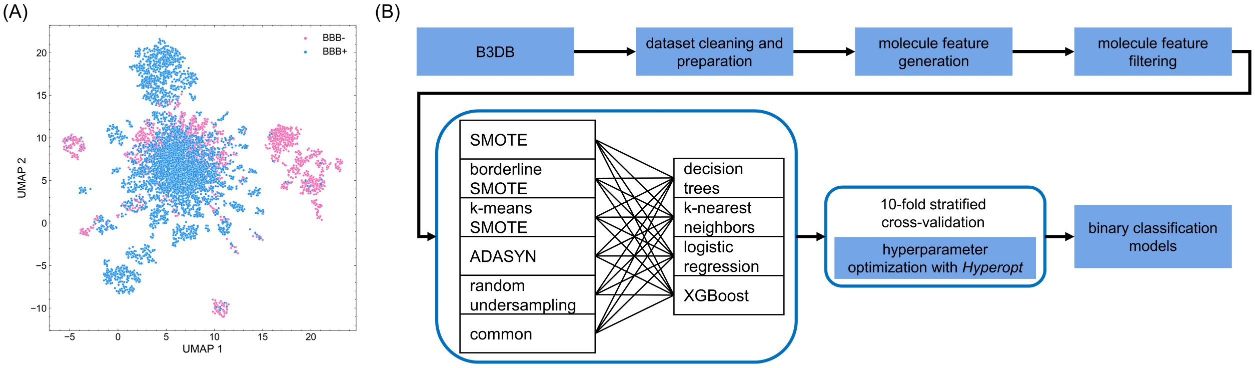
(A)
BBB-
BBB+
20
15
10
5
0
−5
−10
UMAP 2
−5
0
5
10
15
20
UMAP 1
(B)
B3DB
dataset cleaning and preparation
molecule feature generation
molecule feature filtering
SMOTE
borderline SMOTE
k-means SMOTE
ADASYN
random undersampling
common
decision trees
k-nearest neighbors
logistic regression
XGBoost
10-fold stratified cross-validation
hyperparameter optimization with *Hyperopt*
binary classification models

Figure 1 Chemical diversity and computational framework for BBB prediction. (A). Chemical diversity of B3DB dataset with non-linear dimension reduction method, UMAP[52]. (B). General pipeline for constructing classification models for BBB permeability. The linking lines denote combinations of classification algorithms and resampling strategies.

2 Methods and Materials

2.1 Dataset Preparation

The dataset used in this study is the B3DB database[51]. B3DB contains 4956 BBB+ and 2851 BBB- molecules (7807 total). This presents an imbalanced classification problem with an imbalance ratio (IR) of 1.73, where $IR = \frac{N_{\text{majority}}}{N_{\text{minority}}}$[53]. After filtering out charged molecules and structures for which `RDKit` could not generate a valid 3D representation, the dataset was reduced to 4855 BBB+ and 2552 BBB- molecules (7407 total). The dataset preparation workflow is shown in Figure SI. 1 (A).

We used UMAP (nonlinear manifold projection)[52] to visually assess the molecular diversity of B3DB. The difficulty of BBB classification is clear from the extensive overlap between the clouds of BBB+ and BBB- data, as shown previously in Figure 1 (A).

2.2 External Dataset Curation

The external dataset was curated from multiple sources: (1) CNS drugs (BBB+) from *DrugBank* using the WHO Anatomical Therapeutic Chemical (ATC) Classification [1]; (2) H1-antihistamines, including both first- and second-generation H1-antagonists[54–56]; (3) compounds that entered clinical trials but were discontinued due to poor BBB permeability[57–61]. Molecules that overlapped with the training dataset (B3DB)[51], as well as single-atom species, were excluded. After filtering, the external dataset contained 171 BBB+ and 4 BBB- compounds. The SMILES strings and labels of the external dataset are given in Table SI. 1. Geometry optimization and feature generation procedures were consistent with those used for B3DB. This newly curated data has been added to the B3DB dataset[51].

2.3 3D Coordinate Generation and Geometry Optimization

When the PubChem CID is available for a molecule, the 3D coordinates are downloaded from `PubChem` web server using `PubChemPy`[62]; otherwise, the 2D coordinates are retrieved. For database entries without a valid CID, 3D coordinates are generated from the isomeric Simplified Molecular Input Line Entry System (SMILES) using `OpenBabel`[63]. Hydrogen atoms are added if necessary, and molecular geometries are then optimized using the MMFF94s force field[64] as implemented in `OpenBabel`[63]. Default parameters were used, except that we allow up to 10,000 iterations. Molecules for which no satisfactory 3D coordinates could be generated or whose geometry optimization failed were eliminated.

2.4 Molecular Feature Generation

Chemical descriptors were chosen to encode the molecules, not only for their computational efficiency but also motivated by recent studies suggesting their advantage over graph neural networks[65]. We used `PaDEL` to compute 1875 descriptors, including 1D, 2D and 3D features[66]. Five molecules were removed because they were incompatible with `PaDEL`:

[1] Downloaded from https://pubchem.ncbi.nlm.nih.gov/classification

H_2O, CH_4, HN_2O, N_2H_4 and H_4NO_2. As a result, each molecule is represented by 1875 descriptors, of which 431 are 3D features.

2.5 Molecular Feature Selection

To build a robust model, feature selection (also known as variable elimination)[67–69] is essential. A simple pipeline was used for feature selection, as shown in Figure SI. 1 (B). Similar to previous classifiers for BBB permeability, we filtered features based on their numerical and statistical properties[47,48,50,70]. Specifically, features with infinite values ($-\infty$ and ∞), NaN values, or extremely large magnitude ($> 10^5$) were removed from the feature matrix. Constant features were discarded, and highly correlated (or duplicate) features were eliminated using a Pearson correlation threshold of 0.8. This process left 475 features. Eliminating linearly correlated features helps remove redundant information from the feature matrix, as shown in Figure SI. 1 (C) and (D). While the performance gains from this step are not critical for this study, it will be beneficial when applying our model to large-scale database screenings.

2.6 Cross-Validation and Hyperparameter Optimization

The general workflow for building classification models is shown in Figure 1 (B) and includes dataset pre-processing, feature engineering (feature generation and selection), stratified 10-fold cross-validation, and hyperparameter optimization. Given the imbalanced nature of our dataset, we combined 4 basic classification algorithms (decision trees, k-nearest neighbours, logistic regression, and XGBoost) with 4 oversampling methods (SMOTE, Borderline SMOTE, k-means SMOTE, and adaptive synthetic (ADASYN)), 1 undersampling approach (random undersampling). As a control, we also considered the performance of the 4 algorithms without any resampling, denoted as `common`. In total, we constructed $4 \times 6 = 24$ classification models, as indicated in Figure 1. The figures and tables are labeled accordingly. For example, the label "xgb-borderline_SMOTE" refers to a model that uses XBGoost as a base classification algorithm, trained on data oversampled with Borderline_SMOTE.

For each model, a modified stratified 10-fold cross-validation was employed for model selection, which ensures the proportions of BBB+ and BBB- molecules remained fixed. In each fold, 90% data points were held out as the training set and the remaining 10% was split equally into testing and validation subsets in a stratified manner. The training subset was processed with resampling strategies to generate synthetic data, which was then fed into the hyperparameter optimization module together with the validation subset. Hence, each instance of 10-fold cross-validation generates 10 sets of hyperparameters. To identify the optimal hyperparameters, we performed another round of stratified 10-fold cross-validation, evaluating the error for each of the 10 hyperparameter sets. In this second step, 90% of the data was used for training, and the remaining 10% was used to assess the model's performance. The hyperparameter set that resulted in the lowest error was then selected for further analysis, as shown in Figure SI. 2.

The hyperparameters were optimized with `Hyperopt`[71] using the tree of Parzen estimators (TPE)[72,73] algorithm, which is a Sequential Model-Based Global Optimization (SMBO) algorithm.

2.7 Performance Evaluations

We assessed the models' performance with a set of extensive evaluation metrics, which can be categorized into threshold metrics (e.g., accuracy), ranking metrics (e.g., AUC), and probabilistic metrics. Since our focus is on imbalanced classification problems, we primarily used threshold metrics. Specifically, we considered sensitivity (also known as recall, hit rate, or true positive rate), specificity (also known as selectivity or true negative rate), precision (also known as positive predictive value), accuracy, F_1 score, Matthews correlation coefficient (MCC), geometric mean score (GEOM), and balanced accuracy score (BACC), as defined in Equation 1 to Equation 5[74,75]. These evaluation metrics are computed with confusion matrix elements, true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

$$sensitivity = \frac{TP}{TP+FN} \tag{1}$$

$$specificity = \frac{TN}{FP+TN} \tag{2}$$

$$precision = \frac{TP}{TP+FP} \tag{3}$$

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

$$F_1 = 2 \cdot \frac{precision \cdot sensitivity}{precision + sensitivity} \tag{5}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}} \tag{6}$$

$$GEOM = \sqrt{sensitivity \cdot specificity} \tag{7}$$

$$BACC = \frac{\text{sensitivity} + \text{specificity}}{2} \tag{8}$$

We also used the ROC-AUC to explore the relationship between true positives and false positives across different probability thresholds as a ranking metric. The shape of the ROC curve and the area under the curve (AUC) provide useful insights into classifier performance. Random classifiers yield an AUC of 0.5, which serves as a baseline for model ranking. The precision-recall curve demonstrates the relationship between *precision* and *recall* and is more informative than ROC, especially for binary imbalanced problems[76]. The average precision

(AP) score is commonly used in information retrieval[77], defined as

$$AP = \sum_n (R_n - R_{n-1}) P_n \tag{9}$$

where R_n and P_n are the precision and recall at the n-th threshold. Higher values of AP indicate better models.

3 Results and Discussions

3.1 Choice of Classification Algorithm

The performance of all 24 classifiers for 10 groups of hyperparameters is reported in Figure SI. 3 - Figure SI. 6. Notably, all sets of hyperparameters yield similar results, despite being computed using different validation data. Performance across different data samples is also consistent, with small standard deviations (cf. Table SI. 2)), supporting the robustness and generalizability of our models.

We first compare the performance of different classification models in predicting BBB permeability. Cross-validation results show that the statistical variance of the algorithms, from lowest to highest, follows the order: XGBoost < kNN < logistical regression < decision trees (Figure 2, Figure SI. 7 and Figure SI. 8). Similarly, the average ROC_AUC scores follow the same ranking, with XGBoost achieving the highest performance, followed by kNN, logistic regression, and decision trees. Overall, XGBoost shows the highest accuracy and consistency, making it the most reliable model.

3.2 Choice of Sampling Strategy

The impact of sampling strategies was analyzed by comparing them to their counterparts without resampling (i.e., “common” models), as shown in Figure 2, Figure SI. 7 and Figure SI. 9. Undersampling tends to give inferior performance, as shown by lower performance scores. This is unsurprising, since our dataset is relatively small and the imbalance

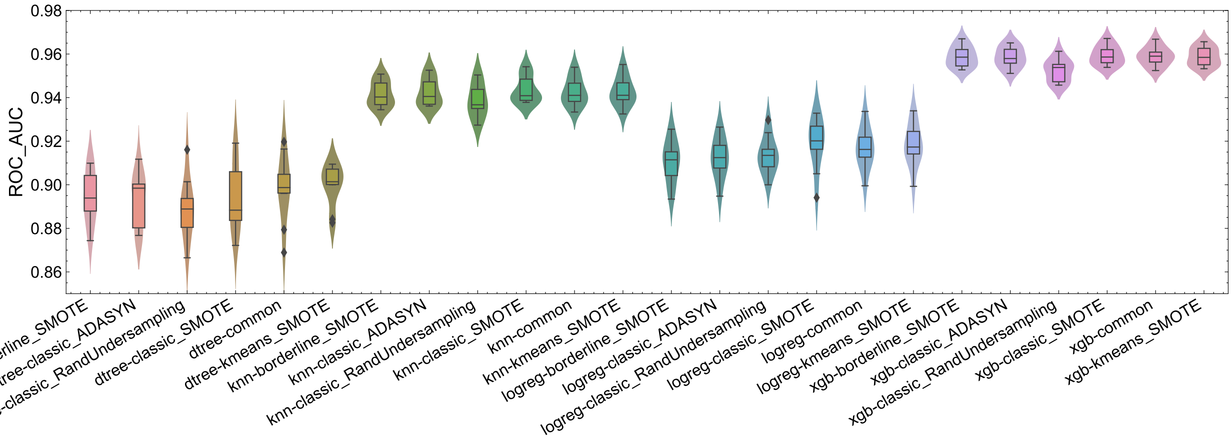

Figure 2 Area under the curve (AUC) for ROC curves of 24 predictive models by combining XGBoost, kNN, logistic regression, and decision trees with various sampling strategies, respectively

between the majority (BBB+) and minority (BBB-) classes is not extreme, so that the loss of information from undersampling outweighs its potential benefits. Although XGBoost with undersampling achieved the highest specificity due to its low false positive rate, this came at the cost of sensitivity, as the model had a tendency to produce false negatives. In addition, undersampling with kNN tends to give similar performance to traditional kNN, probably because kNN captures the pairwise similarity of input molecules and is less sensitive to the effects of undersampling.

In comparison, oversampling improves the model performance as expected (Figure 3). Among the four oversampling methods evaluated, Borderline SMOTE and ADASYN outperform classic SMOTE and k-means SMOTE when applied to decision trees and kNN (Figure SI. 9). k-means SMOTE is the best-performing variant for logistic regression. All oversampling strategies achieve similar performance on XGBoost based models, suggesting that XGBoost does not greatly benefit from oversampling strategies. This is likely because the class imbalance in our dataset is moderate, and also because XGBoost is already a robust ensemble learning method.

A summary of the best-performing resampling strategy for each base ML model is shown in Table 1, and the corresponding ROC_AUC results are in Figure 4. The predictive perfor-

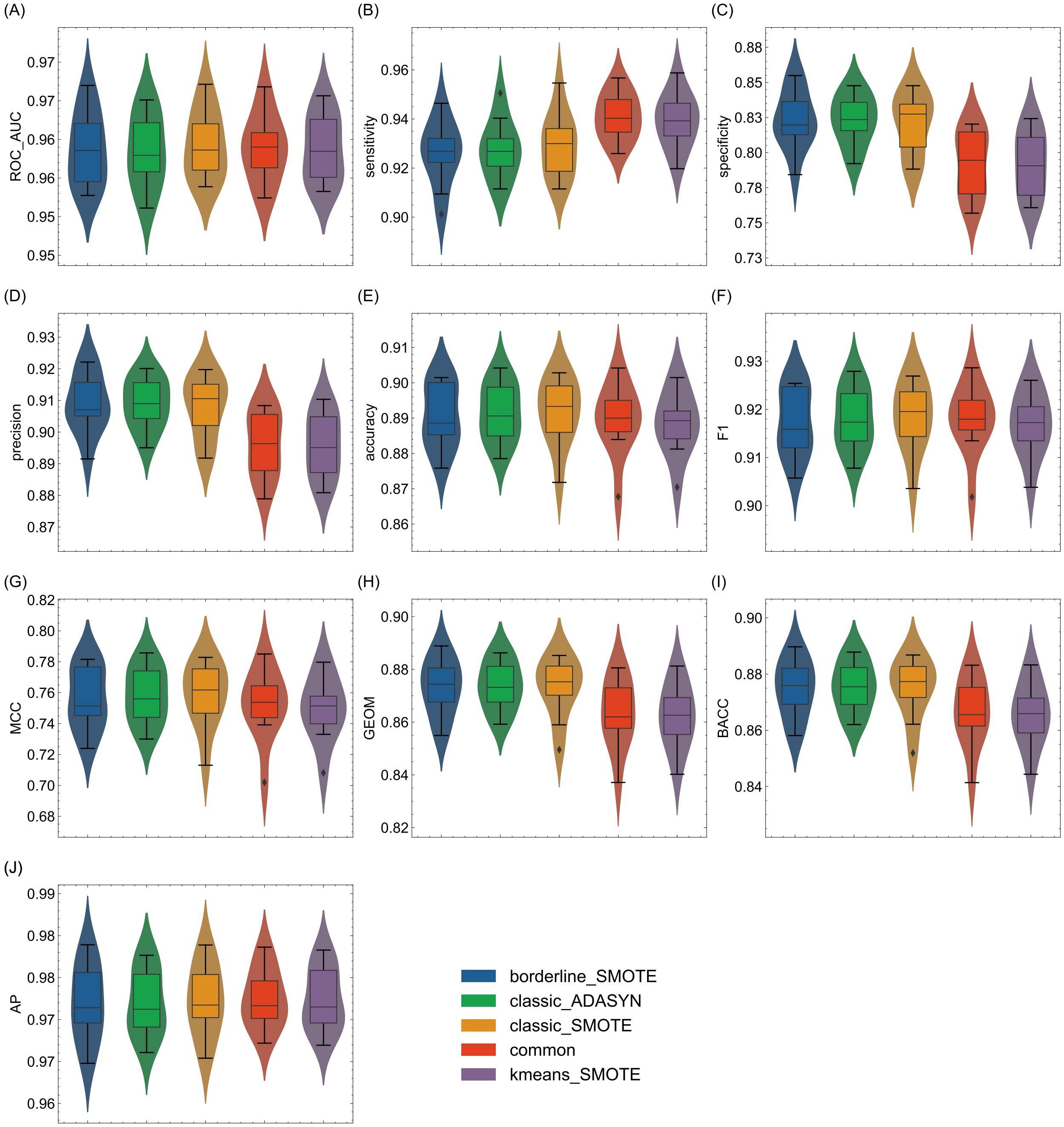

Figure 3 Model performances for top XGBoost models with oversampling strategies, including the raw form of XGBoost (denoted as *common*).

mance of the top models in each category, as demonstrated by the ROC and precision-recall curves, aligns with the model comparisons discussed earlier in Section 3.1.

Table 1 Performance summary of best resampling strategy for each base ML model.

model_name	ROC_AUC	sensitivity	specificity	precision	accuracy	F1	MCC	GEOM	BACC	AP
dtree-kmeans_SMOTE	0.9000 ±0.0094	0.8814 ±0.0158	0.7472 ±0.0289	0.8692 ±0.0120	0.8352 ±0.0101	0.8751 ±0.0079	0.6332 ±0.0227	0.8113 ±0.0136	0.8143 ±0.0124	0.9300 ±0.0100
knn-classic_SMOTE	0.9436 ±0.0062	0.9067 ±0.0098	0.8248 ±0.0237	0.9079 ±0.0114	0.8785 ±0.0112	0.9073 ±0.0084	0.7312 ±0.0254	0.8647 ±0.0141	0.8658 ±0.0136	0.9649 ±0.0063
logreg-classic_SMOTE	0.9190 ±0.0120	0.8622 ±0.0192	0.8225 ±0.0315	0.9026 ±0.0163	0.8485 ±0.0168	0.8818 ±0.0135	0.6730 ±0.0363	0.8419 ±0.0187	0.8424 ±0.0186	0.9491 ±0.0105
xgb-classic_ADASYN	0.9585 ±0.0048	0.9275 ±0.0121	0.8233 ±0.0180	0.9090 ±0.0080	0.8916 ±0.0083	0.9181 ±0.0065	0.7583 ±0.0184	0.8737 ±0.0094	0.8754 ±0.0091	0.9770 ±0.0040

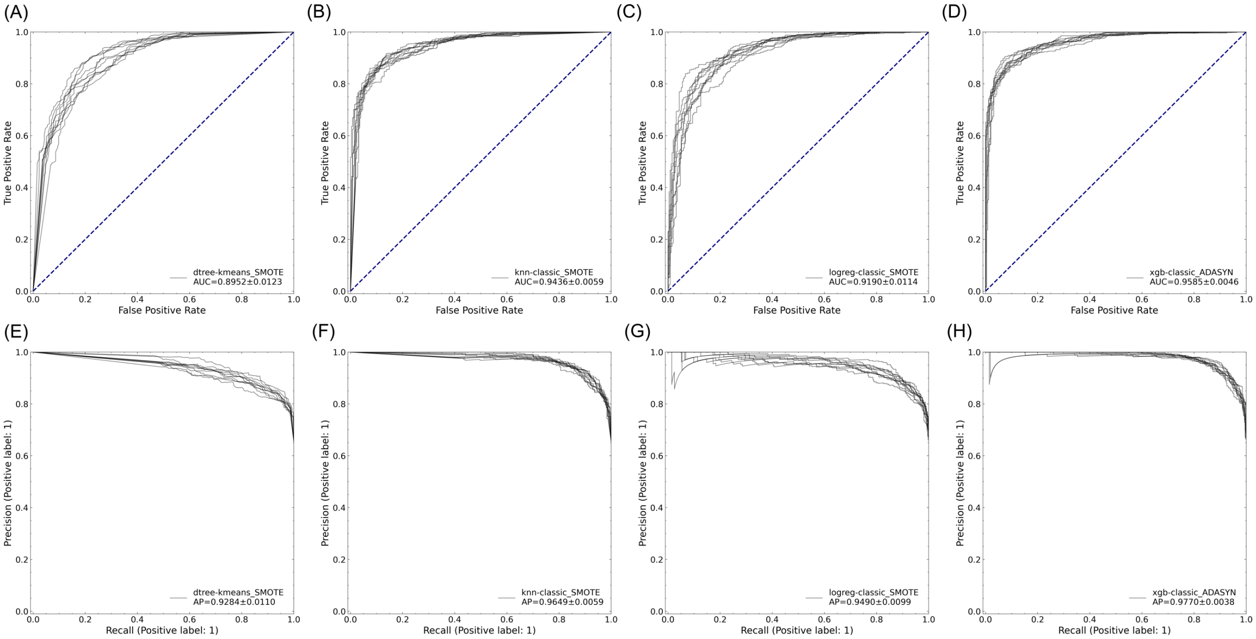

Figure 4 ROC curves and precision-recall curves for each classification algorithm with 10-fold cross-validation.

3.3 Combined Model Performance

Based on a comprehensive analysis of all performance metrics, especially ROC_AUC and AP, we can identify xgb-classic_ADASYN, xgb-borderline_SMOTE and xgb-classic_SMOTE as the top 3 model-sampling strategy combinations, as shown in Figure 2, Figure SI. 7, Figure SI. 10 and Figure SI. 11. The averaged performance metrics of xgb-borderline_SMOTE and xgb-classic_ADASYN are almost identical. However, xgb-classic_ADASYN is more robust, as indicated by slightly smaller variation across the 10-fold splits (cf., the last column in Figure SI. 9). The sensitivity of xgb-kmeans_SMOTE and raw XGBoost is slightly better than that of xgb-classic_ADASYN, xgb-borderline_SMOTE, and xgb-classic_SMOTE, which is caused by the higher rate of false negatives over true positives of xgb-kmeans_SMOTE and raw XGBoost models, as indicated in Figure 3.

The kNN based models give ROC_AUC between 0.9387 and 0.9436. Notably, knn-kmeans_SMOTE (AP = 0.9502) and knn-common (AP = 0.9547) give the highest AP scores, implying their superior performance at picking BBB+ molecules (true positive samples). However, a substantial drop in sensitivity is observed for both knn-borderline_SMOTE model and knn-classical_ADASYN, compared to xgb-classical_ADASYN (sensitivity = 0.9275), suggesting that the former two kNN models are suboptimal for detecting BBB+ molecules. For a more detailed assessment of model performance, please see Figure SI. 7 and Table SI. 2.

3.4 Evaluation Model Generalizability with External Dataset

To evaluate the generalizability of our models, we used the newly curated external evaluation dataset of 175 compounds (171 BBB+ and 4 BBB-) described in Section 2.2. We then compared the performances of our models with six previously published models[78–82]. The performance metrics are summarized in Table 2. We used the Area Under the Precision-Recall Curve (AP) as the primary performance metric, because our prediction dataset was highly imbalanced (BBB permeable to impermeable compounds $>$ 40:1). In such extremely skewed datasets, AP provides a more reliable measure of model performance. Model rankings based on ROC_AUC are generally consistent with AP-based trends, while rankings based on other metrics, such as MCC, may differ. This is likely due to the extreme class imbalance in the prediction dataset, which contains very few BBB- compounds. Unlike AP, MCC is threshold-dependent and more sensitive to skewed class distributions.

Our results show that many of our models outperform existing benchmarks in terms of AP scores, indicating improved predictive performance. Notably, models based on XGBoost and decision trees, when combined with resampling techniques including borderline-SMOTE and ADASYN, achieve the highest AP scores. This may be attributed to the ability of XG-

Boost and decision tree models to handle imbalanced data effectively and capture non-linear relationships. Additionally, resampling methods like borderline-SMOTE and ADASYN improve performance by focusing on minority samples near the decision boundary and those that are more difficult to classify. This is consistent with the general model performance as discussed in Section 3.2 and Section 3.3, which highlight XGBoost's superior performance. Interestingly, one decision tree model also performed well in our benchmarking. This might be because XGBoost is already robust to class imbalance and gains less from resampling, while decision trees benefit more from such techniques.

Comparing the performance of our models to that of other published works, it is important to note that all benchmark models were trained on datasets that included approximately 40–60 of the total 175 compounds used in the prediction dataset. In contrast, our models were trained on B3DB, which does not overlap with the prediction dataset. Therefore, the reported performance of previous models may be inflated due to data leakage, while our evaluation provides a more realistic generalization capability. Despite this disadvantage, our models consistently demonstrate performance that is comparable to, and often exceeds, that of models found in the literature. This underscores our models' exceptional predictive capability.

We also present SHAP analysis results for two of our models, XGBoost-borderline_SMOTE and dtree-borderline_SMOTE, on the prediction dataset, as shown in Figure 5. Descriptors related to the molecule's electronic topological state, hydrogen bonding potential, and autocorrelation consistently rank among the most important. Key descriptors include ETA_Shape_Y, nHAvin, SHsOH, and AATSC6e, suggesting that BBB permeability is strongly influenced by molecular shape, polarizability, and hydrophobicity. This is consistent with the known behavior of blood-brain barrier transport, which favors smaller, less polar molecules capable of diffusing across the barrier[83,84].

Table 2 Comparison of the predictive performance of our models with earlier works. Abbreviations used in this table: AutoML-HSL, Automated Machine Learning using `Hyperopt-sklearn`; DNN, Deep Neural Network; GNN, Graph Neural Network; Mixed DL, Mixed Deep Learning.

Source	**Model**	**AP**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1**	**MCC**	**GEOM**	**BACC**	**ROC_AUC**
Present study	dtree-borderline_SMOTE	0.9997	0.8114	0.8070	1.0000	1.0000	0.8932	0.2954	0.8983	0.9035	0.9934
Present study	xgb-classic_ADASYN	0.9993	0.9086	0.9064	1.0000	1.0000	0.9509	0.4258	0.9521	0.9532	0.9708
Present study	xgb-borderline_SMOTE	0.9991	0.8457	0.8421	1.0000	1.0000	0.9143	0.3296	0.9177	0.9211	0.9605
Present study	xgb-classic_RandUndersampling	0.9991	0.7943	0.7895	1.0000	1.0000	0.8824	0.2810	0.8885	0.8947	0.9635
Present study	knn-classic_RandUndersampling	0.9988	0.8743	0.8772	0.7500	0.9934	0.9317	0.2725	0.8111	0.8136	0.9518
Present study	xgb-classic_SMOTE	0.9988	0.8629	0.8596	1.0000	1.0000	0.9245	0.3504	0.9272	0.9298	0.9488
Swanson et al. (2024)[78]	GNN	0.9988	0.9200	0.9298	0.5000	0.9876	0.9578	0.2368	0.6818	0.7149	0.9518
Present study	knn-classic_SMOTE	0.9986	0.8343	0.8304	1.0000	1.0000	0.9073	0.3173	0.9113	0.9152	0.9444
Present study	knn-common	0.9986	0.9257	0.9357	0.5000	0.9877	0.9610	0.2483	0.6840	0.7178	0.9444
Present study	xgb-common	0.9985	0.8743	0.8713	1.0000	1.0000	0.9313	0.3661	0.9335	0.9357	0.9386
Present study	dtree-common	0.9982	0.8743	0.8772	0.7500	0.9934	0.9317	0.2725	0.8111	0.8136	0.9313
Present study	logreg-classic_RandUndersampling	0.9981	0.8857	0.8889	0.7500	0.9935	0.9383	0.2880	0.8165	0.8194	0.9225
Present study	dtree-kmeans_SMOTE	0.9980	0.8686	0.8713	0.7500	0.9933	0.9283	0.2654	0.8084	0.8107	0.9247
Present study	knn-kmeans_SMOTE	0.9980	0.8914	0.9006	0.5000	0.9872	0.9419	0.1924	0.6710	0.7003	0.9167
Present study	knn-borderline_SMOTE	0.9979	0.7200	0.7135	1.0000	1.0000	0.8328	0.2320	0.8447	0.8567	0.9152
Present study	logreg-classic_ADASYN	0.9979	0.8343	0.8304	1.0000	1.0000	0.9073	0.3173	0.9113	0.9152	0.9137
Present study	xgb-kmeans_SMOTE	0.9979	0.8629	0.8596	1.0000	1.0000	0.9245	0.3504	0.9272	0.9298	0.9137
Present study	knn-classic_ADASYN	0.9977	0.7029	0.6959	1.0000	1.0000	0.8207	0.2230	0.8342	0.8480	0.9079
Present study	logreg-borderline_SMOTE	0.9977	0.8457	0.8421	1.0000	1.0000	0.9143	0.3296	0.9177	0.9211	0.9050
Present study	logreg-common	0.9977	0.8914	0.9006	0.5000	0.9872	0.9419	0.1924	0.6710	0.7003	0.9064
Present study	logreg-kmeans_SMOTE	0.9975	0.8857	0.8947	0.5000	0.9871	0.9387	0.1854	0.6689	0.6974	0.8977
Present study	logreg-classic_SMOTE	0.9968	0.8571	0.8713	0.2500	0.9803	0.9226	0.0537	0.4667	0.5607	0.8713
Parakkal et al. (2022)[79]	Mixed DL (10-fold CV)	0.9964	0.8514	0.8538	0.7500	0.9932	0.9182	0.2461	0.8002	0.8019	0.8757
Parakkal et al. (2022)[79]	Mixed DL	0.9958	0.9200	0.9240	0.7500	0.9937	0.9576	0.3495	0.8325	0.8370	0.8582
Kumar et al. (2022)[80]	DNN	0.9957	0.7829	0.7836	0.7500	0.9926	0.8758	0.1899	0.7666	0.7668	0.8567
Han et al. (2025)[81]	AutoML-HSL	0.9923	0.9200	0.9298	0.5000	0.9876	0.9578	0.2368	0.6818	0.7149	0.8180
Present study	dtree-classic_ADASYN	0.9922	0.6800	0.6784	0.7500	0.9915	0.8056	0.1360	0.7133	0.7142	0.7412
Present study	dtree-classic_RandUndersampling	0.9841	0.7771	0.7778	0.7500	0.9925	0.8721	0.1862	0.7638	0.7639	0.7186
Present study	dtree-classic_SMOTE	0.9820	0.7371	0.7544	0.0000	0.9699	0.8487	-0.0859	0.0000	0.3772	0.4620
Tang et al. (2022)[82]	Multi-model		0.8343	0.8343		1.0000	0.9097	0.0000			

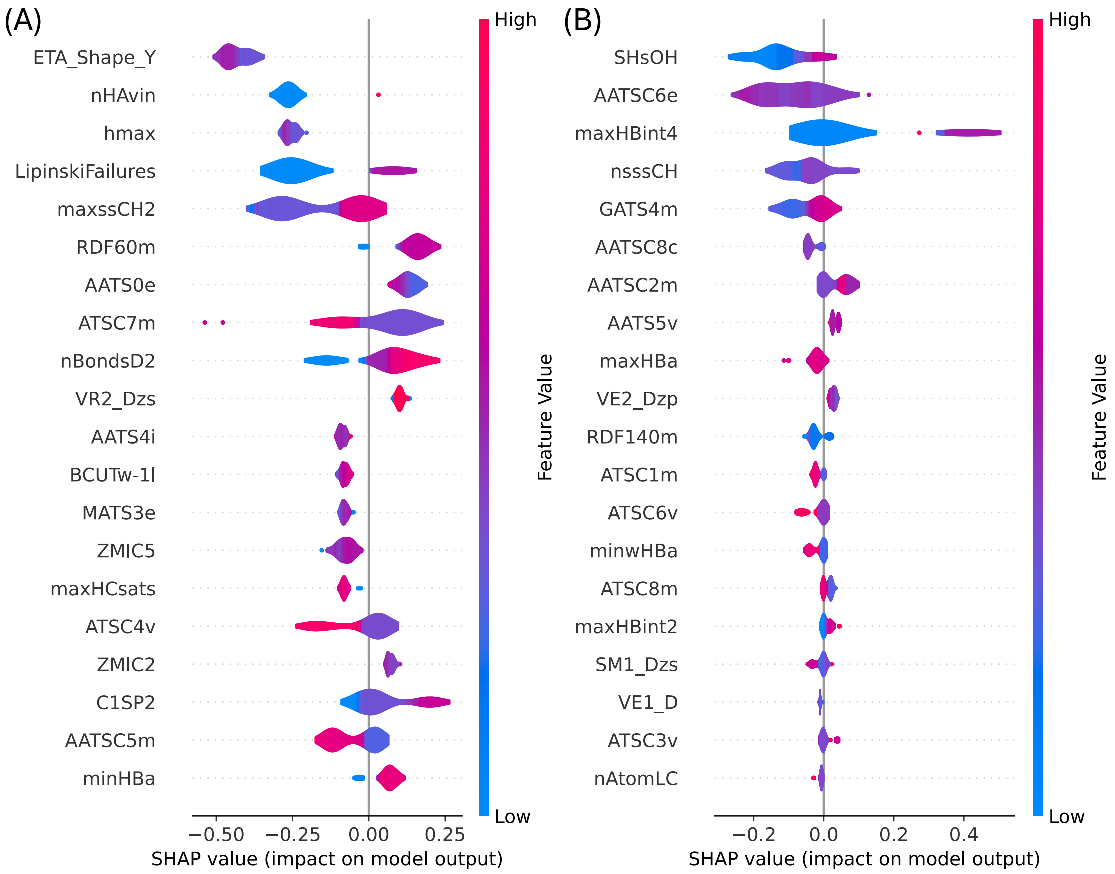

Figure 5 SHAP (SHapley Additive exPlanations) summary plots for (A) XGBoost-borderline_SMOTE and (B) dtree-borderline_SMOTE.

3.5 `B3clf`: An Open-Source Python Package for BBB Predictions

We are disseminating the models proposed in this study as an open-source package, `B3clf`, distributed and hosted by the QC-Devs software consortium[85]. This free and simple command-line tool facilitates early-stage evaluation of molecular permeability in the CNS drug design and development pipeline. It accepts a text file containing SDF filenames or SMILES strings as input. The tool then generates 3D coordinates, optimizes the geometry, and computes molecular descriptors as described in Figure 6. `B3clf` then selects features and applies the pre-trained predictive models to predict whether a molecule is BBB-permeable. The entire workflow can be executed with a single line of `bash` code and typically takes about 4 seconds per molecule (c.f. Figure SI. 12):

```
b3clf -mol input_molecules.sdf -clf xgb -sampling classic_SMOTE -
    out BBB_pred_results.xlsx -verbose 1
```

Predictions generated by `B3clf` are stored in a `CSV` file containing the molecule name, predicted probability of crossing the BBB, and the assigned BBB permeability classification (BBB+ or BBB-). The probability score represents the model's confidence in each prediction and can be directly used to generate ROC and precision–recall curves. This makes `B3clf` a practical and reproducible benchmark for future comparative studies. Moreover, we implemented a web server in Hugging Face (`https://huggingface.co/spaces/QCDevs/b3clf`) with a graphical user interface, lowering the programming requirements for accessing these BBB predictive ML models.

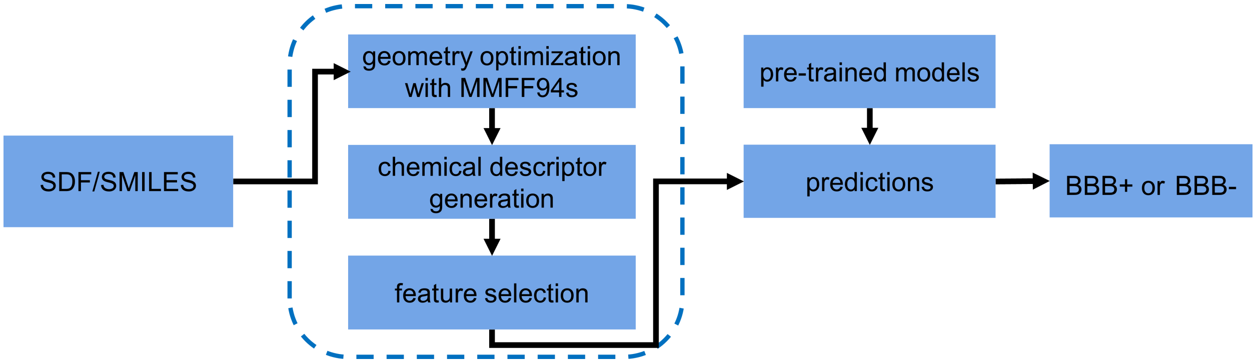

Figure 6 Architecture of `B3clf` computational package.

4 Conclusions

The blood-brain barrier (BBB) functions as a protective boundary for the central nervous system (CNS), maintaining homeostasis and safeguarding it from potentially harmful external agents[1–3]. However, BBB presents significant obstacles in the effective delivery of drug molecules to the CNS, which often leads to failures in clinical trials for CNS drug discovery initiatives[4–6]. Consequently, the prediction of BBB permeability has emerged as a vital consideration in the design and development of CNS-targeted therapeutics. In recent years, predictive models utilizing machine learning (ML) algorithms have attracted considerable attention[12,17–27,35]. However, a persistent challenge remains: class imbalance. This term refers to the unequal distribution of BBB-permeable versus impermeable molecules within training datasets. Such an imbalance may adversely impact the performance of the models and restrict the generalizability of the ML methodologies in this domain.

In this study, we developed ML models by applying resampling strategies to the B3DB dataset[51]. Specifically, we used four classification models (decision trees, kNN, logistic regression, and XGBoost) and six sampling techniques to address the class imbalance problem. The sampling strategies include four oversampling methods (SMOTE, k-means SMOTE, Borderline SMOTE, and ADASYN), one undersampling method (random undersampling), and the original data (no resampling). In total, this yielded 24 models (4 classification algorithms × 6 data configurations). We curated an external dataset of 175 molecules to evaluate the generalizability of our models by using it as an independent test set. We compared our models' predictive performance on this dataset against six other ML models from the literature. The results show that our models consistently achieve comparable or better performance, indicating their strong generalization ability.

To disseminate our methods and ensure the reproducibility of our results, we built a free and open-source Python package `B3clf` whereby users can predict a compound's BBB permeability on the command line or (more advanced) as a Python script. Moreover, we also implemented an easy-to-use web server, hosted by Hugging Face, allowing users to predict the BBB permeability in their browser. Our predictive tool `B3clf` provides both the predictive labels and the corresponding probability, which can be used as a baseline for benchmarking for future BBB predictive models.

While our models provide promising results for BBB permeability predictions, several factors should be considered. Firstly, the reported models may not generalize well to specific experimental measurements of the BBB. This is because (1) the training dataset B3DB does not differentiate between different experimental methods or conditions, and there may be quite limited data for a specific measurement type; (2) the noise in the data should be considered because data were collected from different experiments. Secondly, addressing the noise can help further improve the model's generalizability, which is not covered by this study. Lastly, the threshold probability is not optimized for predicting BBB labels. This is beyond the scope of this study, and one can refer to Ref. [86] for more information.

Data and Software Availability

The B3clf model, including all source code and installation instructions, is available at `https://github.com/theochem/B3clf`. A web interface ready for direct use is freely and publicly accessible at `https://huggingface.co/spaces/QCDevs/b3clf`.

Author Contributions

- Fanwang Meng: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing - review & editing.
- Jitian Chen: Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing - review & editing.
- Juan Samuel Collins-Ramirez: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing - review & editing.
- Paul W. Ayers: Conceptualization, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Writing – original draft, Writing - review & editing.

Conflicts of interest

The authors declared no conflicts of interest.

Acknowledgement

P.W.A. acknowledges Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada Research Chairs, and the Digital Research Alliance of Canada for financial and computational support. F.M. acknowledges the Banting Postdoctoral Fellowship administered by Canada's three research granting agencies: the Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Social Sciences and Humanities Research Council (SSHRC). F.M. also thanks the support of the Alliance's DRI (Digital Research Infrastructure) EDIA (Equity, Diversity, Inclusion and Accessibility) Champions Pilot Program, Digital Research Alliance of Canada, for financial and computational support. F.M. also thanks the support of the Center for Advanced Computing (CAC) at Queen's University. J.C. acknowledges the Canada Graduate Scholarships Master's (CGS-M) program for financial support.

Supporting Information Available

The Supporting Information is available free of charge at https:/xxx/doi/xxx/xxx.xxx.

References

(1) Risau, W.; Wolburg, H. Development of the blood-brain barrier. *Trends in Neurosciences* **1990**, *13*, 174–178.

(2) Profaci, C. P.; Munji, R. N.; Pulido, R. S.; Daneman, R. The blood–brain barrier in health and disease: Important unanswered questions. *Journal of Experimental Medicine* **2020**, *217*.

(3) Daneman, R. The blood–brain barrier in health and disease. *Annals of Neurology* **2012**, *72*, 648–672.

(4) Pardridge, W. M. Blood–brain barrier delivery. *Drug Discovery Today* **2007**, *12*, 54–61.

(5) Gust, J.; Hay, K. A.; Hanafi, L.-A.; Li, D.; Myerson, D.; Gonzalez-Cuyar, L. F.; Yeung, C.; Liles, W. C.; Wurfel, M.; Lopez, J. A.; others Endothelial activation and blood–brain barrier disruption in neurotoxicity after adoptive immunotherapy with CD19 CAR-T cells. *Cancer Discovery* **2017**, *7*, 1404–1419.

(6) Marei, H. E.; Althani, A.; Afifi, N.; Hasan, A.; Caceci, T.; Pozzoli, G.; Cenciarelli, C. Current progress in chimeric antigen receptor T cell therapy for glioblastoma multiforme. *Cancer Medicine* **2021**,

(7) Li, Y.; Chen, T.; Miao, X.; Yi, X.; Wang, X.; Zhao, H.; Lee, S. M.-Y.; Zheng, Y. Zebrafish: A promising in vivo model for assessing the delivery of natural products, fluorescence dyes and drugs across the blood-brain barrier. *Pharmacological Research* **2017**, *125*, 246–257.

(8) Delsing, L.; Herland, A.; Falk, A.; Hicks, R.; Synnergren, J.; Zetterberg, H. Models of the blood-brain barrier using iPSC-derived cells. *Molecular and Cellular Neuroscience* **2020**, 103533.

(9) Jeong, S.; Kim, S.; Buonocore, J.; Park, J.; Welsh, C. J.; Li, J.; Han, A. A three-dimensional arrayed microfluidic blood–brain barrier model with integrated electrical sensor array. *IEEE Transactions on Biomedical Engineering* **2017**, *65*, 431–439.

(10) Mulvihill, J. J.; Cunnane, E. M.; Ross, A. M.; Duskey, J. T.; Tosi, G.; Grabrucker, A. M. Drug delivery across the blood–brain barrier: recent advances in the use of nanocarriers. *Nanomedicine* **2020**, *15*, 205–214.

(11) Liang, Y.; Yoon, J.-Y. In situ sensors for blood-brain barrier (BBB) on a chip. *Sensors and Actuators Reports* **2021**, 100031.

(12) Muehlbacher, M.; Spitzer, G. M.; Liedl, K. R.; Kornhuber, J. Qualitative prediction of blood–brain barrier permeability on a large and refined dataset. *Journal of Computer-Aided Molecular Design* **2011**, *25*, 1095–1106.

(13) Carpenter, T. S.; Kirshner, D. A.; Lau, E. Y.; Wong, S. E.; Nilmeier, J. P.; Lightstone, F. C. A method to predict blood-brain barrier permeability of drug-like compounds using molecular dynamics simulations. *Biophysical Journal* **2014**, *107*, 630–641.

(14) Lombardo, F.; Blake, J. F.; Curatolo, W. J. Computation of brain- blood partitioning of organic solutes via free energy calculations. *Journal of Medicinal Chemistry* **1996**, *39*, 4750–4755.

(15) Thai, N. Q.; Theodorakis, P. E.; Li, M. S. Fast Estimation of the Blood–Brain Barrier Permeability by Pulling a Ligand through a Lipid Membrane. *Journal of Chemical Information and Modeling* **2020**, *60*, 3057–3067.

(16) Wang, Y.; Gallagher, E.; Jorgensen, C.; Troendle, E. P.; Hu, D.; Searson, P. C.; Ulmschneider, M. B. An experimentally validated approach to calculate the blood-brain barrier permeability of small molecules. *Scientific Reports* **2019**, *9*, 1–11.

(17) Gupta, M.; Lee, H. J.; Barden, C. J.; Weaver, D. F. The blood–brain barrier (BBB) score. *Journal of Medicinal Chemistry* **2019**, *62*, 9824–9836.

(18) Xiong, B.; Wang, Y.; Chen, Y.; Xing, S.; Liao, Q.; Chen, Y.; Li, Q.; Li, W.; Sun, H. Strategies for Structural Modification of Small Molecules to Improve Blood–Brain Barrier Penetration: A Recent Perspective. *Journal of Medicinal Chemistry* **2021**,

(19) Zhang, L.; Zhu, H.; Oprea, T. I.; Golbraikh, A.; Tropsha, A. QSAR modeling of the blood–brain barrier permeability for diverse organic compounds. *Pharmaceutical Research* **2008**, *25*, 1902–1914.

(20) Luco, J. M.; Marchevsky, E. QSAR studies on blood-brain barrier permeation. *Current Computer-Aided Drug Design* **2006**, *2*, 31–55.

(21) Wang, W.; Kim, M. T.; Sedykh, A.; Zhu, H. Developing enhanced blood–brain barrier permeability models: integrating external bio-assay data in QSAR modeling. *Pharmaceutical Research* **2015**, *32*, 3055–3065.

(22) Iyer, M.; Mishra, R.; Han, Y.; Hopfinger, A. Predicting blood–brain barrier partitioning of organic molecules using membrane–interaction QSAR analysis. *Pharmaceutical Research* **2002**, *19*, 1611–1621.

(23) Norinder, U.; Haeberlein, M. Computational approaches to the prediction of the blood–brain distribution. *Advanced Drug Delivery Reviews* **2002**, *54*, 291–313.

(24) Garg, P.; Verma, J. In silico prediction of blood brain barrier permeability: an artificial neural network model. *Journal of Chemical Information and Modeling* **2006**, *46*, 289–297.

(25) Alsenan, S.; Al-Turaiki, I.; Hafez, A. A deep learning approach to predict blood-brain barrier permeability. *PeerJ Computer Science* **2021**, *7*, e515.

(26) Achiaa Atwereboannah, A.; Wu, W.-P.; Nanor, E. Prediction of Drug Permeability to the Blood-Brain Barrier using Deep Learning. 4th International Conference on Biometric Engineering and Applications. 2021; pp 104–109.

(27) Kortagere, S.; Chekmarev, D.; Welsh, W. J.; Ekins, S. New predictive models for blood–brain barrier permeability of drug-like molecules. *Pharmaceutical Research* **2008**, *25*, 1836–1845.

(28) Singh, M.; Divakaran, R.; Konda, L. S. K.; Kristam, R. A classification model for blood brain barrier penetration. *Journal of Molecular Graphics and Modelling* **2020**, *96*, 107516.

(29) Liu, L.; Zhang, L.; Feng, H.; Li, S.; Liu, M.; Zhao, J.; Liu, H. Prediction of the Blood–Brain Barrier (BBB) Permeability of Chemicals Based on Machine-Learning and Ensemble Methods. *Chemical Research in Toxicology* **2021**,

(30) Deconinck, E.; Zhang, M. H.; Coomans, D.; Vander Heyden, Y. Classification tree models for the prediction of blood- brain barrier passage of drugs. *Journal of Chemical Information and Modeling* **2006**, *46*, 1410–1419.

(31) Zhao, Y. H.; Abraham, M. H.; Ibrahim, A.; Fish, P. V.; Cole, S.; Lewis, M. L.; de Groot, M. J.; Reynolds, D. P. Predicting penetration across the blood-brain barrier from simple descriptors and fragmentation schemes. *Journal of Chemical Information and Modeling* **2007**, *47*, 170–175.

(32) Suenderhauf, C.; Hammann, F.; Huwyler, J. Computational prediction of blood-brain barrier permeability using decision tree induction. *Molecules* **2012**, *17*, 10429–10445.

(33) Yuan, Y.; Zheng, F.; Zhan, C.-G. Improved prediction of blood–brain barrier permeability through machine learning with combined use of molecular property-based descriptors and fingerprints. *The AAPS Journal* **2018**, *20*, 1–10.

(34) Roy, D.; Hinge, V. K.; Kovalenko, A. To pass or not to pass: predicting the blood–brain barrier permeability with the 3D-RISM-KH molecular solvation theory. *ACS Omega* **2019**, *4*, 16774–16780.

(35) Martins, I. F.; Teixeira, A. L.; Pinheiro, L.; Falcao, A. O. A Bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of Chemical Information and Modeling* **2012**, *52*, 1686–1697.

(36) Rebala, G.; Ravi, A.; Churiwala, S. *An Introduction to Machine Learning*; Springer International Publishing, 2019; pp 57–66.

(37) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **2002**, *16*, 321–357.

(38) More, A. Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv preprint arXiv:1608.06048* **2016**,

(39) Estabrooks, A.; Jo, T.; Japkowicz, N. A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence* **2004**, *20*, 18–36.

(40) Drummond, C.; Holte, R. C.; others C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. Proceedings of the International Conference on Machine Learning (ICML 2003) Workshop on Learning from Imbalanced Data Sets II. 2003; pp 1–8.

(41) Fernández, A.; Garcia, S.; Herrera, F.; Chawla, N. V. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research* **2018**, *61*, 863–905.

(42) Han, H.; Wang, W.-Y.; Mao, B.-H. Borderline-SMOTE: a new over-sampling method

in imbalanced data sets learning. International Conference on Intelligent Computing. 2005; pp 878–887.

(43) Last, F.; Douzas, G.; Bacao, F. Oversampling for imbalanced learning based on k-means and smote. *arXiv preprint arXiv:1711.00837* **2017**,

(44) He, H.; Bai, Y.; Garcia, E. A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). 2008; pp 1322–1328.

(45) Bunkhumpornpat, C.; Sinapiromsaran, K.; Lursinsap, C. DBSMOTE: density-based synthetic minority over-sampling technique. *Applied Intelligence* **2012**, *36*, 664–684.

(46) Pereira, T.; Abbasi, M.; Oliveira, J. L.; Ribeiro, B.; Arrais, J. Optimizing blood–brain barrier permeation through deep reinforcement learning for *de novo* drug design. *Bioinformatics* **2021**, *37*, i84–i92.

(47) Wang, Z.; Yang, H.; Wu, Z.; Wang, T.; Li, W.; Tang, Y.; Liu, G. In silico prediction of blood–brain barrier permeability of compounds by machine learning and resampling methods. *ChemMedChem* **2018**, *13*, 2189–2201.

(48) Alsenan, S.; Al-Turaiki, I.; Hafez, A. A Recurrent Neural Network model to predict blood–brain barrier permeability. *Computational Biology and Chemistry* **2020**, *89*, 107377.

(49) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016; pp 785–794.

(50) Shi, Z.; Chu, Y.; Zhang, Y.; Wang, Y.; Wei, D. Prediction of blood-brain barrier permeability of compounds by fusing resampling strategies and eXtreme gradient boosting. *IEEE Access* **2020**, *9*, 9557–9566.

(51) Meng, F.; Xi, Y.; Huang, J.; Ayers, P. W. A curated diverse molecular database of blood-brain barrier permeability with chemical descriptors. *Scientific Data* **2021**, *8*.

(52) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* **2018**,

(53) Noorhalim, N.; Ali, A.; Shamsuddin, S. M. Handling imbalanced ratio for class imbalance problem using smote. Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017). 2019; pp 19–30.

(54) Simon, F. E. R.; Simons, K. J. H1 antihistamines: current status and future directions. *World Allergy Organization Journal* **2008**, *1*, 145–155.

(55) Tamai, I.; Kido, Y.; Yamashita, J.; Sai, Y.; Tsuji, A. Blood-brain barrier transport of H1-antagonist ebastine and its metabolite carebastine. *Journal of Drug Targeting* **2000**, *8*, 383–393.

(56) YAMAZAKI, M.; FUKUOKA, H.; NAGATA, O.; KATo, H.; ITO, Y.; TERASAKI, T.; TSUJI, A. Transport mechanism of an H1-antagonist at the blood-brain barrier: transport mechanism of mepyramine using the carotid injection technique. *Biological and Pharmaceutical Bulletin* **1994**, *17*, 676–679.

(57) Imbimbo, B. P. Why did tarenflurbil fail in Alzheimer's disease? *Journal of Alzheimer's Disease* **2009**, *17*, 757–760.

(58) Sacco, R. L.; DeRosa, J. T.; Haley Jr, E. C.; Levin, B.; Ordronneau, P.; Phillips, S. J.; Rundek, T.; Snipes, R. G.; Thompson, J. L.; Investigators, G. A.; others Glycine antagonist in neuroprotection for patients with acute stroke: GAIN Americas: a randomized controlled trial. *Jama* **2001**, *285*, 1719–1728.

(59) Alavijeh, M. S.; Chishty, M.; Qaiser, M. Z.; Palmer, A. M. Drug metabolism and

pharmacokinetics, the blood-brain barrier, and central nervous system drug discovery. *NeuroRx* **2005**, *2*, 554–571.

(60) Ahlskog, J. E. Common myths and misconceptions that sidetrack Parkinson disease treatment, to the detriment of patients. Mayo Clinic Proceedings. 2020; pp 2225–2234.

(61) Fellner, S.; Bauer, B.; Miller, D. S.; Schaffrik, M.; Fankhänel, M.; Spruß, T.; Bernhardt, G.; Graeff, C.; Färber, L.; Gschaidmeier, H.; others Transport of paclitaxel (Taxol) across the blood-brain barrier in vitro and in vivo. *The Journal of clinical investigation* **2002**, *110*, 1309–1318.

(62) Swain, M. PubChemPy. 2017; https://github.com/mcs07/PubChemPy.

(63) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **2011**, *3*, 1–14.

(64) Halgren, T. A. MMFF VI. MMFF94s option for energy minimization studies. *Journal of Computational Chemistry* **1999**, *20*, 720–729.

(65) Jiang, D.; Wu, Z.; Hsieh, C.-Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *Journal of Cheminformatics* **2021**, *13*, 1–23.

(66) Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry* **2011**, *32*, 1466–1474.

(67) Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R. P.; Tang, J.; Liu, H. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)* **2017**, *50*, 1–45.

(68) Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Computers & Electrical Engineering* **2014**, *40*, 16–28.

(69) Dash, M.; Liu, H. Feature selection for classification. *Intelligent Data Analysis* **1997**, *1*, 131–156.

(70) Shaker, B.; Yu, M.-S.; Song, J. S.; Ahn, S.; Ryu, J. Y.; Oh, K.-S.; Na, D. LightBBB: computational prediction model of blood–brain-barrier penetration based on LightGBM. *Bioinformatics* **2021**, *37*, 1135–1139.

(71) Bergstra, J.; Komer, B.; Eliasmith, C.; Yamins, D.; Cox, D. D. Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science & Discovery* **2015**, *8*, 014008.

(72) Bergstra, J.; Bardenet, R.; Bengio, Y.; Kégl, B. Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems* **2011**, *24*.

(73) Bergstra, J.; Yamins, D.; Cox, D. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. International Conference on Machine Learning. 2013; pp 115–123.

(74) Japkowicz, N. *Imbalanced Learning*; John Wiley & Sons, Ltd, 2013; Chapter 8, pp 187–206.

(75) He, H.; Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* **2009**, *21*, 1263–1284.

(76) Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS ONE* **2015**, *10*, e0118432.

(77) Schütze, H.; Manning, C. D.; Raghavan, P. *Introduction to Information Retrieval*; Cambridge University Press Cambridge, 2008; Vol. 39.

(78) Swanson, K.; Walther, P.; Leitz, J.; Mukherjee, S.; Wu, J. C.; Shivnaraine, R. V.;

Zou, J. ADMET-AI: a machine learning ADMET platform for evaluation of large-scale chemical libraries. *Bioinformatics* **2024**, *40*, btae416.

(79) Cherian Parakkal, S.; Datta, R.; Das, D. DeepBBBP: high accuracy blood-brain-barrier permeability prediction with a mixed deep learning model. *Molecular informatics* **2022**, *41*, 2100315.

(80) Kumar, R.; Sharma, A.; Alexiou, A.; Bilgrami, A. L.; Kamal, M. A.; Ashraf, G. M. DeePred-BBB: A blood brain barrier permeability prediction model with improved accuracy. *Frontiers in neuroscience* **2022**, *16*, 858126.

(81) Han, H.; Shaker, B.; Lee, J. H.; Choi, S.; Yoon, S.; Singh, M.; Basith, S.; Cui, M.; Ahn, S.; An, J.; others Employing Automated Machine Learning (AutoML) Methods to Facilitate the In Silico ADMET Properties Prediction. *Journal of Chemical Information and Modeling* **2025**, *65*, 3215–3225.

(82) Tang, Q.; Nie, F.; Zhao, Q.; Chen, W. A merged molecular representation deep learning method for blood–brain barrier permeability prediction. *Briefings in Bioinformatics* **2022**, *23*, bbac357.

(83) Kadry, H.; Noorani, B.; Cucullo, L. A blood–brain barrier overview on structure, function, impairment, and biomarkers of integrity. *Fluids and Barriers of the CNS* **2020**, *17*, 69.

(84) Faramarzi, S.; Kim, M. T.; Volpe, D. A.; Cross, K. P.; Chakravarti, S.; Stavitskaya, L. Development of QSAR models to predict blood-brain barrier permeability. *Frontiers in Pharmacology* **2022**, *13*, 1040838.

(85) Chan, M.; Verstraelen, T.; Tehrani, A.; Richer, M.; Yang, X. D.; Kim, T. D.; Vöhringer-Martinez, E.; Heidar-Zadeh, F.; Ayers, P. W. The Tale of HORTON: Lessons Learned in a Decade of Scientific Software Development. *The Journal of Chemical Physics* **2024**, *160*, 162501.

(86) Garg, A.; Ramamurthi, N.; Das, S. S. Addressing Imbalanced Classification Problems in Drug Discovery and Development Using Random Forest, Support Vector Machine, AutoGluon-Tabular, and H2O AutoML. *Journal of Chemical Information and Modeling* **2025**, *65*, 3976–3989.